Exploiting User Profiles to Support Differentiated Services in Next-Generation Wireless Networks

Vijoy Pandey[1], Dipak Ghosal[2], and Biswanath Mukherjee[2]

vijoy@nortelnetworks.com, ghosal@cs.ucdavis.edu, mukherje@cs.ucdavis.edu

[1]Nortel Networks
4655 Great America Parkway
Santa Clara, CA 95054, U.S.A.

[2]Department of Computer Science
University of California
Davis, CA 95616, U.S.A.

Correspondence author: Dipak Ghosal, *ghosal@cs.ucdavis.edu*,
Tel: +1-530-754-9251, Fax: +1-530-752-4767

July 15, 2002

Abstract

In the next-generation wireless network, user profiles such as the location, the velocity (both speed and direction), and the resource requirements of the mobile device can be accurately determined and maintained by the network on a per-user basis. We investigate the design of a differentiated-services architecture which exploits user profiles to maximize the network efficiency and which supports differentiated service classes, each with different Quality-of-Service (QoS) guarantees. In this paper, we provide implementation guidelines of such an architecture for the Third-Generation Partnership Project (3GPP) network. The key underlying primitive of the architecture is the use of user profiles to perform advance resource reservation in target cells of the wireless cellular network. We identify the design tradeoffs and present performance results for an architecture consisting of two service classes, namely (1) a higher-cost profiled service with higher QoS, and (2) a lower-cost non-profiled service with best-effort QoS. Our analysis indicates that a significant decrease in the dropping probability[1] — and, hence, higher QoS — can be guaranteed to users who subscribe to the profiled service. We examine the tradeoffs associated with some of the key system parameters including the reservation distance and the reservation granularity, and we determine their values which maximize the improvement in the dropping probability for all users.

Keywords and Phrases: Wireless network, Quality of service, Differentiated services, User profile, Dropping probability, Call handoff control, Resource-reservation algorithm.

---

[1] Dropping probability is the probability that an admitted call fails due to an unsuccessful handoff.

1 Introduction

The next-generation wireless network [1, 2, 3] will support a rich set of multimedia applications similar to those available in wired networks. To achieve the goal of providing high-quality multimedia services to anyone, anywhere, and at any time [4], network designers will need to implement new techniques that can support Quality of Service (QoS) while accounting for limited bandwidth and for the delay and error characteristics of the wireless network [5]. To support and guarantee QoS, the next-generation wireless network must implement a differentiated-services architecture. This architecture would contain multiple service levels, each with a different QoS guarantee.

Implementing a differentiated-services architecture in a wireless network is complex due to both changes in the resource requirements of the mobile device and the user mobility. Resource requirements may change depending upon changes in the mobile device. For example, a user switching to her laptop from her palmtop might request higher bandwidth from the network. User mobility requires that the network guarantee resources spatially as the user moves across the wireless network.

User mobility in wireless networks has been characterized by random-walk or Brownian-motion-based mathematical models [6, 7, 8, 9]. In reality, however, user mobility has a much higher degree of predictability due to temporal and spatial locality. Temporal locality refers to the fact that a mobile user typically takes predictable routes, which implies that a user will typically cross the same set of cells at predictable times in a wireless cellular network. For instance, a mobile user will typically follow the same path to work in the morning, and the reverse route back home in the evening. Spatial locality refers to the fact that user mobility is constrained along pathways and highways which results in a mobile user crossing the cells in an ordered sequence determined by the manner in which these pathways intersect the cellular coverage area.

Recently, with the FCC E-911 mandate, location-based services have become top priority for cellular service providers. The technology strives to deliver personal, time-critical, location-dependent information to the user, such as route directions, current traffic conditions, tracking other family members, and local facilities-based services. Due to this growing market, a number of different methods to accurately determine the location of a mobile user have evolved. Advances have also been made in velocity-estimation algorithms [10]. Moreover, due to the strong interplay between the different layers — physical, data link, network, and applications layers — of a wireless network, it is possible to accurately determine the resource requirements of the mobile device.

The real-time and aggregate values of a user's mobility and resource requirements are known as the "user profile". The goal of our study is to investigate the design and implementation of a user-profile-based differentiated-services architecture for the next-generation wireless network.

Specific implementation details have been provided for the Third-Generation Partnership Project (3GPP) [3] cellular network architecture, but should translate easily to most other wireless architectures as well. We have studied the performance benefits of our proposed approach in a network with two types of users - (1) profiled users who subscribe to a higher-cost profiled service which guarantees higher QoS and (2) regular (non-profiled) users who receive best-effort service. We observe that the network provides improved QoS to profiled users by significantly reducing their dropping probability through advanced reservation of cell resources along the path predicted by the user profile. There are optimal values of the reservation distance (which is the distance prior to a cell crossing when the reservation is attempted) and the reservation granularity (which is related to the frequency of the re-attempts when a reservation attempt fails) which result in the maximal improvement in dropping probability. Finally, we compare our approach with the static two-class reservation scheme proposed in [11, 12], showing that our approach can achieve improved dropping probability performance.

The remainder of this paper is organized as follows. Section 2 describes our profile-based cellular network architecture and our channel-reservation algorithms. Section 3 outlines the concept of a User Profile Register (UPR) and discusses its design. Section 4 introduces our user-profile-based resource-management algorithm. Implementation details for the 3GPP network architecture have been provided in Section 5. Section 6 describes our simulation model and Section 7 illustrates our numerical results. Section 8 discusses related work in the field. Finally, Section 9 concludes the paper.

2 Existing and Emerging Network Infrastructure

Figure 1(a) shows the components of a cellular network architecture[2] [3]. The Base Transceiver Station (BTS) provides radio coverage to one cell. In the 3GPP architecture, the BTS is a specific instance of a component known as Node B. The Base Station Controller (BSC) manages one or more cells, and hence BTSs. A BSC is responsible for channel assignment and management of handoffs between neighboring BTSs within its area of control. The Mobile Switching Center (MSC) constitutes the interface between the radio system and the fixed network. It is responsible for (1) setting up, managing, and clearing calls (connections); (2) routing incoming calls to the appropriate cell; and (3) managing inter-BSC handoffs. The Gateway MSC (GMSC) performs routing of calls to another MSC area, i.e., it manages inter-MSC handoffs.

---

[2] We have chosen the 3GPP network architecture for our examples, but the concepts outlined in this investigation should translate easily to other wireless cellular architectures as well.

Mobility management is supported by location registers. These registers contain information on the location of a mobile user. The Home Location Register (HLR) is a database containing user-specific data — such as user identity, subscribed services, and current user location — for all users registered in a particular geographical area. The Visitor Location Register (VLR) contains information about all users currently "visiting" its particular geographical area. The Authentication Center (AuC) and the Equipment Identity Register (EIR) are additional databases which aid in identification and authentication of mobile users.

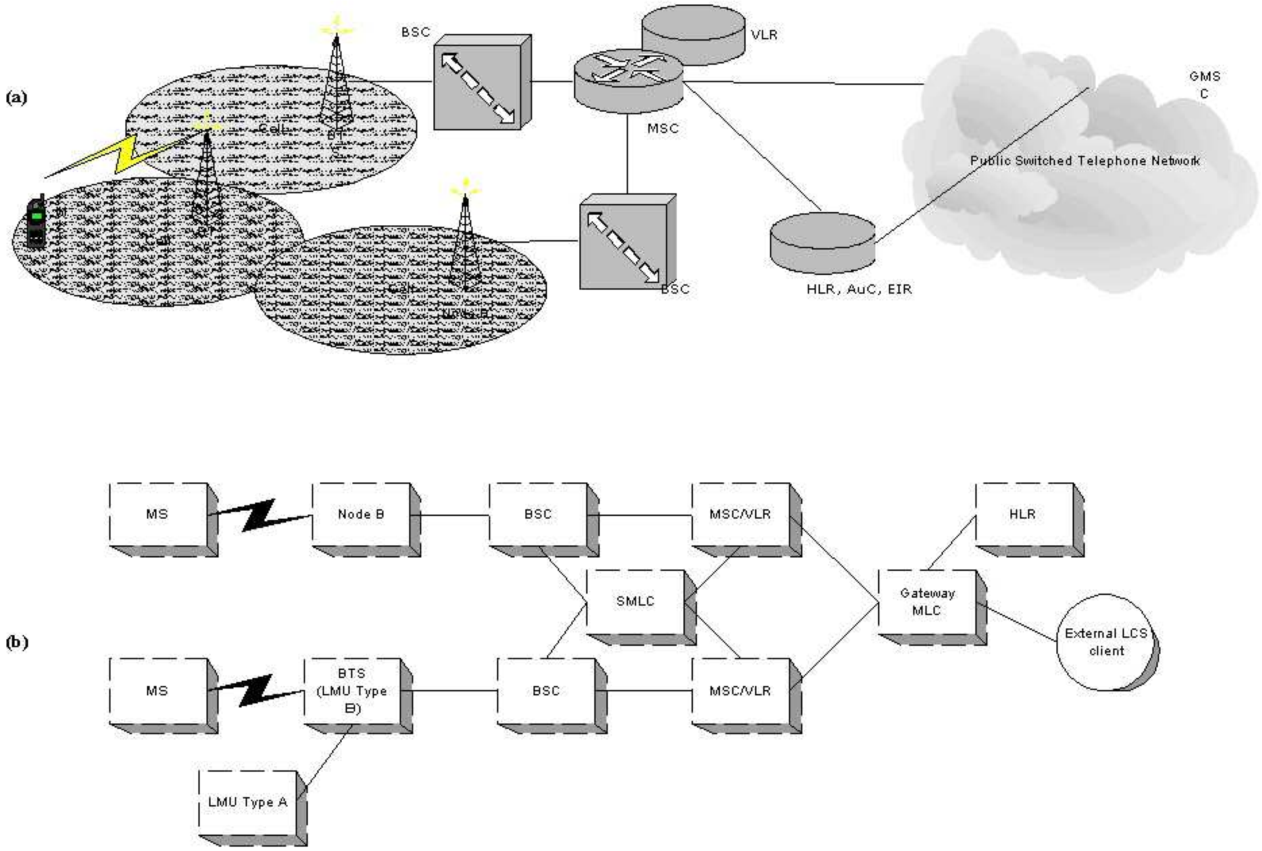

Figure 1: (a) Components of the 3GPP network architecture. (b) Network elements needed to support profile-based channel reservation.

2.1 Infrastructure for LoCation Services

Wireless service providers are rapidly shifting focus from simple voice services to mobile LoCation Services (LCS) [13, 14] which utilize a user's position information to provide localized and personalized services. LCS is dependent on accurate positioning technologies which strive to measure the geographical location of a mobile user as she moves through a wireless cellular network. Support for location measurement is still nascent in current wireless networks. Since location information is a key component of the user profile, we briefly discuss the emerging LCS architecture below.

Figure 1(b) shows the *logical* network elements required to implement location measurement in a 3GPP network. All of these logical elements are implemented within the physical network elements shown in Figure 1(a).

- *Gateway Mobile Location Center (GMLC):* The GMLC is responsible for interfacing with the external world, i.e., with the LCS clients who request the mobile's position. The GMLC is responsible for authenticating the LCS client, requesting the mobile's position with the desired accuracy from the network, and performing any format conversion between the client and the network.

- *Serving Mobile Location Center (SMLC):* The SMLC determines the geographical coordinates of the mobile (and the potential error) in accordance with the quality requested from the GMLC and the capability of the mobile. This logical element can be implemented as a separate physical element or within the BSC.

- *Location Measurement Unit (LMU):* The LMU's role is to help the SMLC take synchronization measurements. LMUs can be integrated within Node B elements such as BTSs (Type B LMUs) or can be distributed over the network (Type A LMUs). Its key function is to determine the location and covert it into some meaningful coordinates X,Y. There are different methods of doing this. (1) Cell-ID and Enhanced Cell-ID schemes locate the nearest tower, and deliver the location of the mobile within a few kilometers. (2) E-OTD/TOA (Enhanced Observed Time Difference/Time of Arrival) uses special transmitters, receivers, and RF signals (above and beyond those used to transmit the calls) to determine the distance of a device from several towers. These distances can be used to determine the location of the mobile. (3) Global Positioning System (GPS) uses receivers integrated into the mobile device to communicate with the GPS satellite constellation, and provides a highly accurate measurement of the mobile's coordinates. (4) Finally, there are software approaches to location measurement based on the signal-strength information from a number of base stations.

In addition, the LCS architecture proposals suggest modifications to a number of existing network elements [13, 14]. For external LCS clients, the MSC should be capable of verifying their authenticity with help from the HLR database, in a manner similar to user verification done today. The MSC should also be able to verify the capability of the mobile in delivering its position with the required accuracy. The BSC should be modified to implement the SMLC functionality, or to control a number of physically-independent SMLCs. Finally, the Node Bs or BTSs have to be modified to possibly include LMU functionality.

3 User Profile and User Profile Register

We introduce an architectural component called the **User Profile Register (UPR)** which is a database similar to HLR and VLR. The UPR contains user-profile information which can be queried for by the wireless network to provide differentiated services to its customers. A user profile consists of mobility patterns and services accessed by the mobile user, tabulated against the time of the day and the day of the week. It contains pointers to network elements which can provide real-time values of the user's location and velocity information. A UPR should have interfaces to external information sources — such as network information databases, described later — to aid in QoS management. The components of the UPR are outlined below.

- **User Location Interface:** This is a logical interface to devices such as the SMLC and the LMUs which can provide real-time user location values to the UPR.
- **User Velocity Interface:** This logical interface will query the network element responsible for real-time velocity estimation of a mobile user [10, 15].
- **User Path Table (UPT):** This table is an ordered list of the most probable paths a mobile user could traverse at any given time on any day of the week. A mobile path is a list of Cell-IDs which a mobile user traverses. For example, let $< c_1, c_2, ..., c_n >$ represent a mobile user's path which starts from cell c_1 and ends in cell c_n.

  This path could contain a number of *hot-spots*. A hot-spot is defined as a collection of cells within a geographical region where the mobile-user population density is very high, e.g., greater than a pre-defined threshold. City downtown regions, train stations, airports, or residential areas represent typical hot-spots. Hot-spots could be dynamic in nature, and can change depending upon the time of the day, traffic conditions and special events. A city downtown can be a hot-spot during the work day, while residential neighborhoods can become hot-spots on weekends. The network can detect a hot-spot near a stadium on Superbowl Sunday, while a snow storm could cause multiple hot-spots along the highways near Lake Tahoe.

  The user could make a call and terminate it at any point along this path, $< c_1, c_2, ..., c_n >$. If all the cells in a mobile user's path are contained within a single hot-spot, the path is considered *internal* to the hot-spot, and does not appear in the UPT.

A User Path Table can include two types of paths:

- *User-specified:* These paths are defined by the user before usage. For example, a user may define a few *typical* paths she might take during an average workday, when signing up for the cellular service. Paths can also be defined using mapping services such as MapQuest or using GPS devices by specifying the end points of the route. The pre-calculated route can then be "pushed" out to the UPR.
- *Statistical:* The UPT could be initialized to an empty value or to the user-specified list above. Subsequently, every time a path is traversed by the user, its ranking is increased in the UPT. Over time, the UPT would typically converge to a fixed set of paths, as users statistically take the same routes for the same chores.

- **User Resource Table (URT):** This table is an ordered list of resources (services) a mobile user uses at any given time on any day of the week. The URT contains two types of entries:
  - *User-specified:* The user could supply a list of services to be used for each path that she could take. For example, voice services could be accessed during the drive to work and back, while web services could be accessed during the lunch break.
  - *Statistical:* The URT could be initialized to an empty value or to the user-specified list above, and subsequently updated by a positive weight every time that service is accessed along the mobile path. The URT will also be constantly updated, and it would converge over time to a virtually static list.
- **Interfaces to External Information Systems:** This entry contains variables obtained from external information systems such as network architecture databases or GPS and Global Information Services (GIS) devices. For example, network architecture databases could supply information about cell layout and sizes. GPS and GIS devices could supply real-time changes in hot-spot definitions since, as discussed earlier, hot-spots could be dynamic in nature and can change depending upon the time of the day, traffic conditions, and special events.

3.1 An Example

Consider the Greater Sacramento geographical region shown in Figure 2. We have highlighted the trajectory of a mobile user traveling from the Arden Town suburb to Richards Boulevard, near downtown Sacramento, to work. She would drive on Howe Avenue to US-50, follow US-50 West, drive on I-5 North, and finally take the Richards Boulevard offramp, near downtown Sacramento.

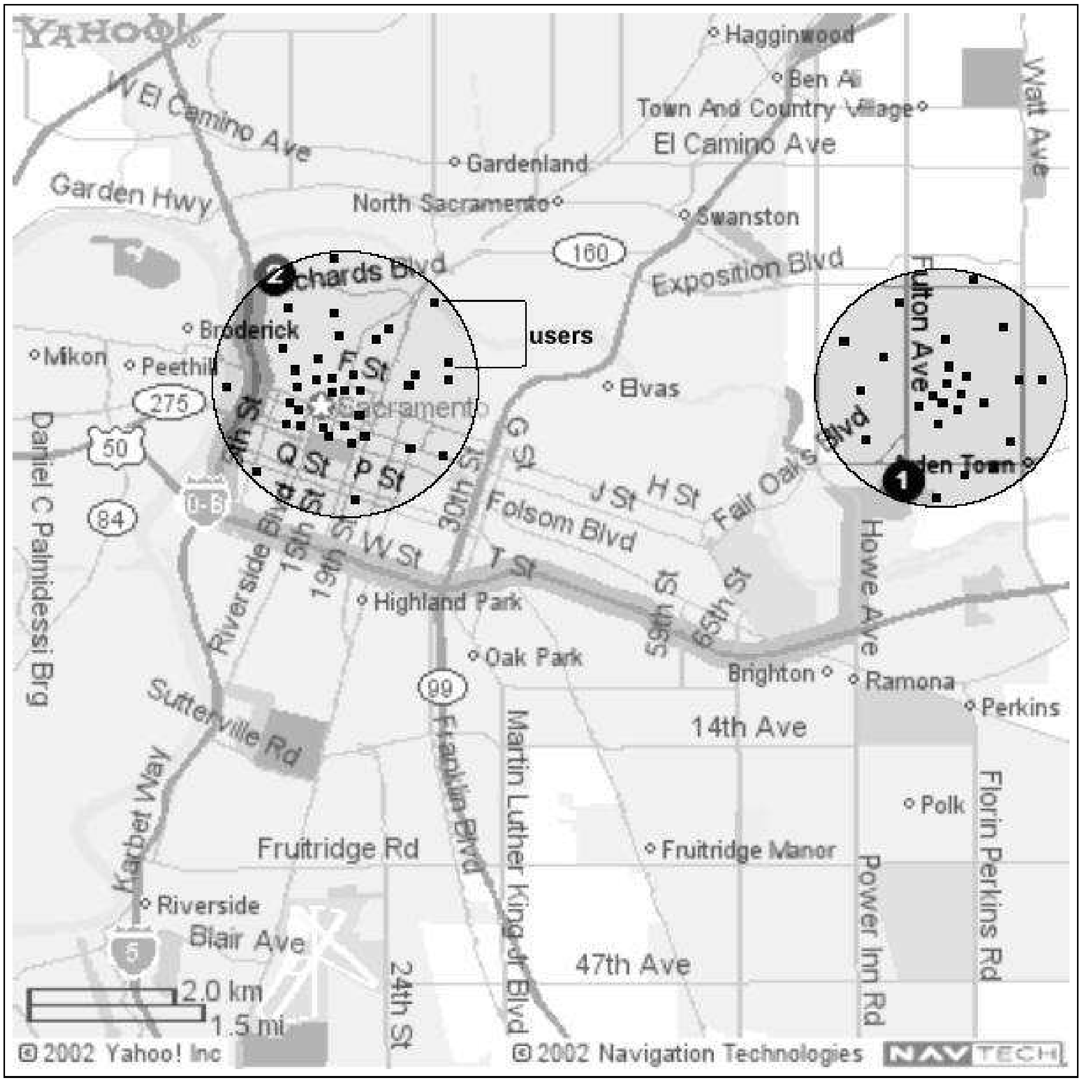

Figure 2: User trajectory and hot-spots in the Sacramento metropolitan area.

On most days, the user would use this trajectory in the morning. On certain occasions, perhaps because of traffic congestion, she might take a separate route to work. While driving to work, she might use voice services, while on a lunch break she might be more interested in data services, e.g., for catching up on personal email.

Initially, when the mobile user starts the use of wireless services, the UPT and the URT are either empty or initialized by the user as described earlier. Every time the mobile traverses a new path which does not exist in the UPT, or accesses a service along a mobile path that is absent from the URT, an entry is added into these tables. Whenever the mobile repeats a particular path or service, the weight of the entry is increased. The table is then re-ordered in descending order with the largest-weight entry on top. Over time, a user profile would develop for this user. Her UPT would converge to show the highlighted route as the most probable path for weekday mornings. The UPR would reflect her preference for voice during the commute period and data during lunch time in her URT. To keep storage requirements for the profile tables small, less probable paths and services could be phased out over time. The network could employ a multi-level virtual-memory-like storage mechanism and caching schemes such as Least Recently Used (LRU) to keep a subset of the table space in faster memory while phasing out the rest of the information to permanent storage.

4 User-Profile-Based Differentiated-Services Architecture

A differentiated-services architecture based on user profiles would support different classes of service, each providing different QoS. In order to implement such an architecture in a wireless network, the following two key components should be designed, namely (1) *a call-admission-control algorithm* and (2) *a call QoS control framework*. A call-admission-control algorithm attempts to regulate the number of users who are admitted into the network. This control operation merits a separate in-depth investigation, and could be performed based on network load [16], pricing models, or classes of service.

In this work, we mainly focus on a call QoS control framework, which is a set of algorithms and policies that perform resource management in the cellular network and that attempt to guarantee the required QoS to the users admitted into the network. User-profile-based call QoS control algorithms utilize the statistical profiles of the mobile user to manage and control the network resources in order to guarantee the negotiated QoS.

For this study, we have considered two classes of service, for two types of users — (1) profiled users who subscribe to the profiling service, expect better QoS, and hence pay more, and (2) non-profiled users who pay less and expect a "best-effort" service from the network. Below, we outline our user-profile-based resource-management algorithm.

4.1 PARMA: Profile-Assisted Resource-Management Algorithm

Figure 3 shows a part of the trajectory of a user commuting from the Arden Town suburb to Richards Boulevard, near downtown Sacramento. The steps a cellular network would take to reserve resources for a profiled user are outlined below; these steps form our profile-assisted resource-management algorithm (PARMA).

1. When a mobile user is *close* to a cell boundary, the network would consult its subscriber database to find out whether the user is a profiled customer.

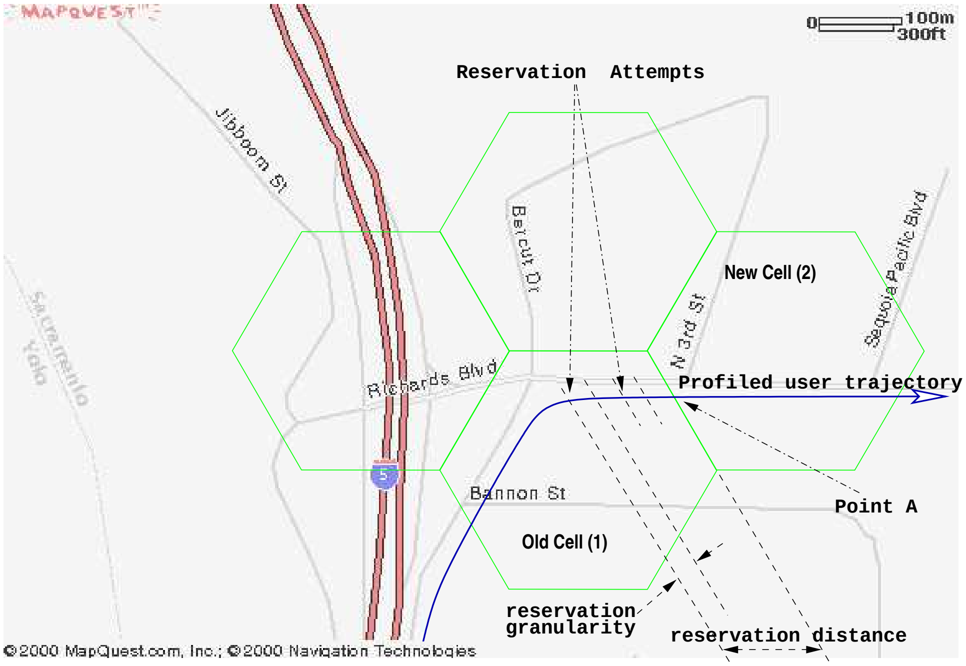

Figure 3: Illustration of PARMA.

2. If the user does not subscribe to the profiling service, then she travels to Point A and attempts a handoff to Cell 2. If there are no channels available in Cell 2 at this handoff instant, the user gets dropped.

3. If the user is a profiled customer, the network consults the UPR and extracts the user's resource requirements from the URT. For the purpose of this discussion, let us assume that the top-most entry in the URT for this part of the mobile's path is voice services, and hence would simply need a channel reservation in the target cell.

4. The network tries to predict the target cell, based on the UPT and the current location and velocity of the mobile.

5. The network would then attempt to reserve a channel in advance for the user in Cell 2 when the user is at a distance r_d from the cell boundary, where r_d is known as the *reservation distance*.

6. If the reservation attempt succeeds, the user is handed off to Cell 2 on the reserved channel at Point A. If the reservation fails, the network re-attempts the reservation every r_g (*reservation granularity*) distance apart, till the reservation succeeds or till a handoff takes place at Point A.

7. If the reservation is unsuccessful till Point A (even after several attempts), then the user session gets dropped. By allowing multiple reservation attempts, the dropping probability of a profiled user can be substantially reduced.

As mentioned in Step 1, PARMA is initiated when the mobile user is *close* to a cell boundary. There are two key approaches to proximity evaluation, and both approaches are conceptually equivalent. In the first approach, the mobile device measures the signal strengths from adjacent base stations and reports this information to the BTS. The network could then decide, based on these measurements, whether the user is "close enough" for PARMA initiation. In the second approach, the network could keep track of the velocity (speed and direction) of a mobile user. Knowing the trajectory of the user from the user profile, the network can then calculate when handoff would occur. Therefore, based on this calculation, the network can decide when to initiate PARMA. In our study, we choose the first approach to measure a user's proximity to a cell boundary.

Though we have considered channel reservation to highlight our algorithm, PARMA is much broader in scope. Any user-specific resource, as defined in the URT for the profiled user, can be allocated for in the target cell. Resources such as the browser cache [17] for mobile browsers, session and state information for data connections, application proxy states for thin clients running on mobile devices, etc., can all be reserved in the target cell through PARMA.

5 Design and Implementation Issues

In this section, we provide some guidelines to implement PARMA in a 3GPP network architecture. First, we describe the modifications we need to perform to the existing network components to support PARMA, and then we propose a possible implementation.

5.1 Modifications to 3GPP Network Elements

We can leverage the location-measurement infrastructure described in Section 2 for obtaining current updates to the position of a profiled user. Furthermore, we have to modify some of the network elements to support PARMA.

- The SMLC should be modified to store a short history of the mobile's position instead of storing just the current position. The size of this history depends on the accuracy of the path-prediction algorithm.

- The UPR database should be implemented to include the profile tables and real-time values for each mobile user in her home area. Each UPR should also be able to accept and incorporate updates to user profiles available from the MSC.

- Logical interfaces should exist between the UPR and external information databases and systems, such as the HLR, network architecture databases, the LMUs, and the SMLC. The HLR interface would aid the UPR in gathering subscription information about a user. Network architecture databases would provide information on hot-spot definitions and cellular layouts. LMUs and the SMLC would provide current location information and would assist in user-velocity estimation.

- The software in the MSC should be enhanced to statistically update the UPT and the URT as described in Section 3. The MSC should be able to make corrections to the user profile depending upon the success or failure of the user-profile-based path-prediction process and the services which the user has accessed, by increasing the "rank" of successfully predicted paths and resource requirements in the UPT and URT, respectively. The MSC should then provide this feedback to the UPR database.

5.2 Implementation of PARMA

Figure 4 shows the message sequence chart for PARMA. In current networks, a mobile device periodically sends a list of BTSs and their signal strengths to the current BTS for the purpose of a handoff, using a *SignalStrengthList* message. In Global System for Mobile Communications (GSM) networks, a mobile device usually sends a list of the 16 strongest BTSs to the current BTS. The current BTS, with help from the MSC, uses this information to decide the target handoff cell. We can modify the BTS software to trigger a *PredictTargetCell* signal to the MSC whenever the signal strength of another BTS comes within a trigger threshold (S_T) of the signal from the current BTS. We study the impact of this threshold on network performance in Section 7.

On receiving the *PredictTargetCell* signal, the MSC sends out a *GetLocation* message to the SMLC requesting the past few coordinates (positional history) of the mobile user. The SMLC functionality can be modified to keep track of the positional history of a mobile user. The length of this history can vary depending on the accuracy of the path-prediction algorithm. On receipt of *GetLocation*, the SMLC forwards the positional history of the mobile to the MSC. For performing path prediction, the MSC also requires the user's path profile (UPT) and the current velocity of the user which are stored in the UPR database. It also requires the URT for gauging the resource requirements of the mobile in the target cell. The *GetProfile* and *SendProfileList* messages accomplish this task.

On receiving the current location, velocity, and the UPT, the MSC performs path prediction and

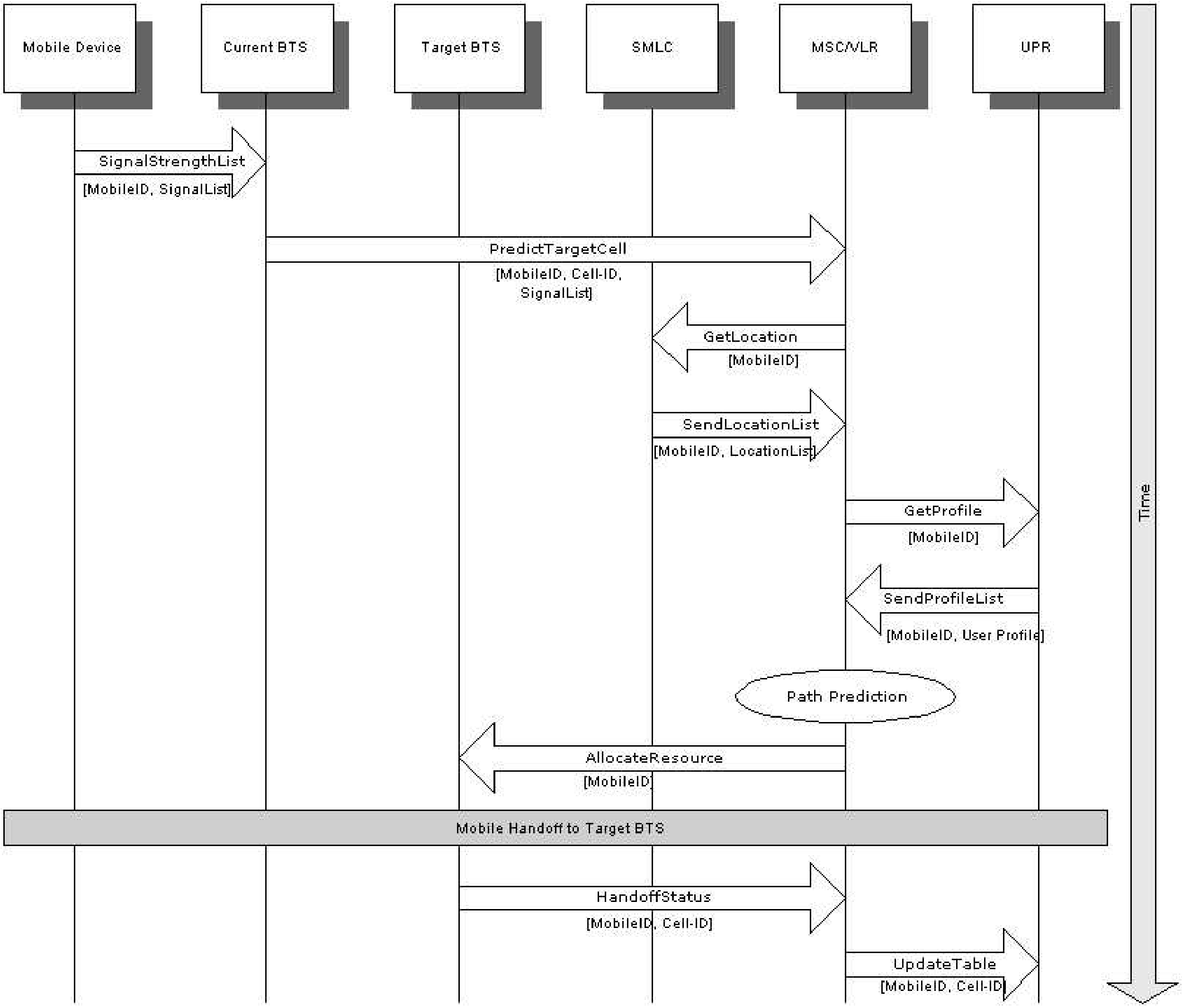

Figure 4: Message Sequence Chart for Profile-Assisted Resource Management Algorithm.

informs the most probable target BTS to reserve resources for the mobile depending on the most probable services accessed by the customer. If the trigger threshold, S_T, is designed accurately, the handoff will occur immediately after the *AllocateResource* message, and the transition to the new cell will be smooth. After the mobile undergoes the handoff, the new BTS sends a status update to the MSC through the *HandoffStatus* signal and includes its own Cell-ID. The MSC checks the Cell-ID to confirm whether this was the cell to which it had sent the *AllocateResource* message. *HandoffStatus* also mentions whether the resource reservation was sufficient or whether the BTS had to allocate fewer or more resources. Finally, the MSC updates the UPT and the URT at the UPR based on the results of the comparison in the previous step.

5.3 Design Issues

There are several issues to consider when implementing PARMA in a wireless network. In this section, we outline some design issues, and we also generalize on the basic scheme presented earlier.

1. *Speed-Modulated Reservation Granularity:* When calculating reservation re-attempts to the target cell, PARMA could take the velocity measurements into account to dynamically modify the reservation granularity. The speed at which a user is traveling towards the target cell would determine how small or large the reservation granularity should be.

2. *Direction-Modulated Reservation Granularity:* The direction of movement of a profiled user with respect to the cell boundaries also plays an important role in the implementation of PARMA. Consider a user who is moving almost tangentially to the cell boundary. The mobile device's list of strongest BTS signals could change very frequently over time, hence triggering very frequent reservation attempts. Therefore, by using direction information coupled with the network's information of cell boundaries, PARMA can employ techniques such as hysteresis and signal thresholds — which are used to avoid the ping-pong effect in handoffs [18] — to reduce unwanted reservation re-attempts.

3. *Hot-spot Modulated Reservation:* The number of permutations of highways and freeways which a user can take to travel between a pair of hot-spots is generally expected to be quite small. Therefore, over time, inter hot-spot portions of a UPT entry tend to converge to a few fixed paths. Hence, path prediction can be performed on inter hot-spot paths with sufficient accuracy by consulting the UPT entries for the user. Moreover, fast table lookups will reduce overheads for resource reservation.

   Intra hot-spot paths are more difficult to predict since the number of permutations of streets that can be taken to reach the destination could be significantly larger as compared to the inter hot-spot scenario. Once the network has determined that the user has entered an intra hot-spot region of her mobile path, the network should solicit help from velocity-estimation algorithms and location-measurement technologies to better predict the user's target cell. The positional history coupled with the current speed and direction measurements will provide better rules for selecting the most probable path in the UPT and, hence, the most probable target cell. This modification will cause some extra overhead at the MSC, but will provide better estimation of the target cell and, therefore result in less resource wastage.

4. *Macrocell Reservation:* Hot-spots are regions of high mobile-user population density. For such geographical areas, the concept of a microcellular/macrocellular multi-tier architecture has

been proposed [19, 20, 21, 22] to provide greater channel capacity through microcells while providing better management using the overlayed macrocells.

If all resource-reservation attempts are unsuccessful, a handoff takes place; and if there are no available resources in the target microcell, the overlayed macrocell can be used to temporarily "hold" the session (and resources) of a profiled user. Under these circumstances, PARMA would keep re-attempting the reservation requests in future target microcells. The goal would be to vertically handoff the session to a microcell at the first available opportunity.

5. *Reservation Granularity versus Round Robin:* PARMA introduces the concept of reservation granularity. Resource-reservation requests are *re-attempted* after every r_g distance till the request succeeds, or a handoff takes place resulting in success or a dropout.

   Instead of the above strategy, all resource-reservation requests could be queued in the target cell if they were unsuccessful on the first attempt. The network could then use a round-robin scheme to scan through the queue and re-attempt each request. The request would get dequeued on a reservation success, or on a successful or unsuccessful handoff.

6. *Signaling Overheads:* Though we have shown nine messages for implementing PARMA (see Figure 4), the overheads are minimal as most messages can be piggy-backed on existing signals, as described below. To aid in handoffs, a mobile device periodically sends the list of closest BTSs to its current BTS. The current BTS can use the same message, coupled with a threshold, S_T, to trigger the *PredictTargetCell* message. *GetLocation* and *SendLocationList* will soon be implemented in the 3G networks to provide LCS. *GetProfile* and *SendProfileList* can be piggy-backed onto HLR database queries for authentication and authorization most of the time (if UPRs are implemented as a part of the HLR database). Moreover, user profiles can be cached in the VLR database inside the MSC and updated at the UPR once the mobile leaves that location area. The *AllocateResource* signal is sent to a target BTS after a handoff decision has been made. For PARMA, we send the signal just after path prediction, imposing no extra overheads. Similarly, the *HandoffStatus* signal is sent irrespective of whether PARMA has been implemented.

   The additional signals in PARMA include the *PredictTargetCell* and *UpdateTable* messages. Moreover, there is some overhead in executing the path-prediction algorithm at the MSC. Hence, the path-prediction algorithm should be lightweight.

6 A Quantitative Analysis

We have studied the performance of PARMA using detailed simulation experiments and an approximate analytical model. Even though we have tried to create an accurate simulation model, we had to make a few deviations from the practical implementation to better analyze the problem. The following section describes our simulation model and how it relates to the practical implementation.

6.1 Simulation Model

We have simulated a single-tier cellular network architecture. The primary resource accessed by cellular users in this network are channels. There are a limited total number of channels, C, available to the network which are allocated to cells using fixed-channel allocation with a static reuse pattern [23]. The reuse distance ratio, i.e., the ratio of the cell radius to the distance from the center of the cell to the next co-channel cell, is denoted as R.

We employ a hexagonal cell structure with a cell radius of c km. We model a limited user population of U users at any given time in the network, out of which a fraction p of the users are profiled. To model the density of users within hot-spots, we use the following algorithm, assuming a circular hot-spot with a radius of H:

1. Choose (or define) the hot-spot radius, H.
2. Choose an angle (δ) as a uniform random number between 0 and 2π.
3. Once δ is fixed, choose a uniform random number ρ between 0 and H. Place a user at an angle δ and at a distance of ρ from the center.
4. Repeat steps 2-3 for each user in a hot-spot.

When this algorithm is used repeatedly, it results in the user density shown in Figure 2, which closely approximates the characteristics of a hot-spot. There are H hot-spots in the region covered by the network. The size and location of these hot-spots can be defined using a graphical interface (which we have developed) to the simulation.

Traditionally, user mobility in wireless and cellular networks has been modeled using simple Brownian-motion or random-walk approximations [21, 24]. For our investigation, we assume that each user has a different direction of movement (D) and speed (V). In reality, a user would have varying speeds and directions over her trajectory depending on the constraints of her terrain and the configuration of the highways pathways. For the purpose of this study, we assume that we can accurately predict a user's trajectory at every given point in time. Hence, modeling each user with a fixed direction of movement does not affect our simulation results. V is chosen uniformly between a specified speed range.

For simulation purposes, PARMA uses the concept of reservation distance to reserve channels for profiled users as described in Section 4.1. In effect, we require a threshold to trigger the channel-reservation algorithm when a profiled user travels close to the edge of a cell. In practice, a threshold on the received signal strength can be used for triggering the algorithm, as described in Section 5.2. Therefore, all the results we have obtained with respect to reservation distance in hexagonal cells are applicable in practice using signal-strength thresholds in arbitrary shaped cells.

We assume that new-calls arrivals into the network follow a Poisson distribution with parameter λ calls/sec. Call-holding time is assumed to follow an exponential distribution with a mean of $1/\mu$ seconds.

6.2 Performance Metrics

- *Total blocking probability*, P_b, is defined as the ratio of the total number of calls blocked or dropped to the total number of new-call attempts made. The total number of calls blocked is the sum of the total number of new-calls blocked and the total number of handoff calls dropped.

- *Dropping probability*, P_d, is the ratio of the number of calls dropped due to a failed handoff to the number of calls which successfully entered the network.

- *Improvement in Dropping Probability*, γ, is the reduction in dropping probability of a profiled user as compared to a non-profiled user. If P_{dp} is the dropping probability for profiled users and P_{dn} is the dropping probability for non-profiled users, then γ is defined as:

$$\gamma = \frac{P_{dn} - P_{dp}}{P_{dn}} * 100\% \tag{1}$$

6.3 Default Parameter Values

Unless otherwise mentioned, the default values for the various parameters are shown in Table 1. Other operating parameters will be introduced when necessary.

7 Results and Discussion

Figure 5 shows the dropping probability for profiled and non-profiled users as a function of the call-arrival rate for $C = 400$ channels. As expected, dropping probability for profiled users is significantly lower than that for non-profiled users. Figure 6 shows the improvement in dropping

Table 1: The default values of the various parameters.

Parameter	Description	Value
H	Number of hot-spots	3
H	Hot-spot radius	3 km
A	Number of cells in region	20×20
c	Cell radius	0.5 km
C	Total number of channels	400
R	Reuse distance ratio	2
V	User speed (range)	25-40 mph
U	Number of users	10,000
p	Fraction of profiled users	0.4
r_d	Reservation distance	0.2 km
r_g	Reservation granularity	0.01 km
$1/\mu$	Mean call-holding time	120 sec
λ	New-call arrival rate	0.025 calls/sec

probability, γ, experienced by profiled users as the new-call arrival rate is increased from 0.005 calls/sec to 0.05 calls/sec for $C = 300$ and $C = 400$ channels.

Let us examine the improvement in dropping probability shown in Figure 6 in relation to the absolute values of the dropping probabilities of profiled and non-profiled users shown in Figure 5. At the left extreme of both figures, when λ is small (e.g., when $\lambda = 0.005$ calls/sec), the load to the network is very light. Hence, very few handoff attempts of both profiled and non-profiled users get dropped, as observed in Figure 5. Therefore, reserving channels for profiled users in advance does not result in a large γ. At the other extreme, when λ is large (e.g., when $\lambda = 0.05$ calls/sec in this example), the network load is high. A significant number of the channel-reservation requests for profiled users get blocked. This causes the dropping probability for both non-profiled and profiled users to be close to each other, resulting in a small γ. When the network load is moderate, the profiled users obtain the most benefit from channel reservation. As can be observed in Figure 5, the dropping probability for non-profiled users keeps increasing even at moderate load, while the dropping probability for profiled users starts to flatten out, benefiting from the reservations. This results in a substantial improvement in dropping probability, with a peak occurring at $\lambda = 0.025$ calls/sec (for $C = 400$ channels), when we observe an improvement of 37%. For $C = 300$ channels,

γ peaks at 34% for $\lambda = 0.02$ calls/sec.

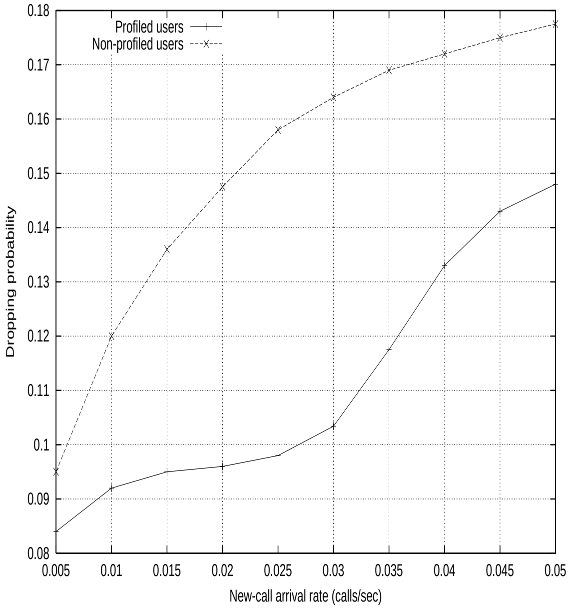

Figure 5: The dropping probability for different new-call arrival rates at C=400 channels.

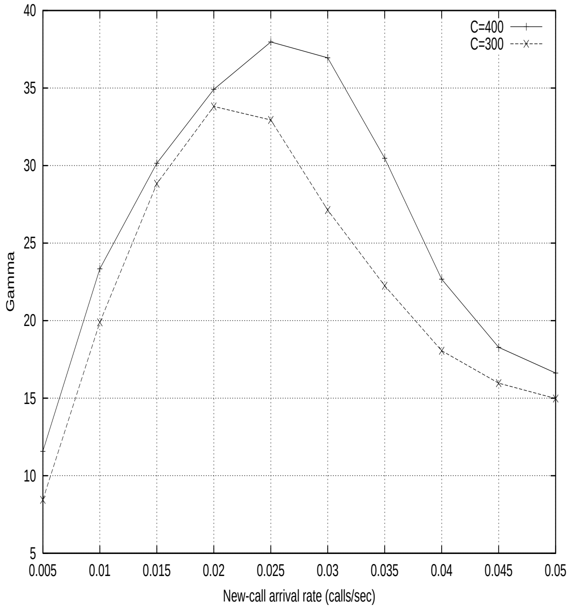

Figure 6: γ for different new-call arrival rates.

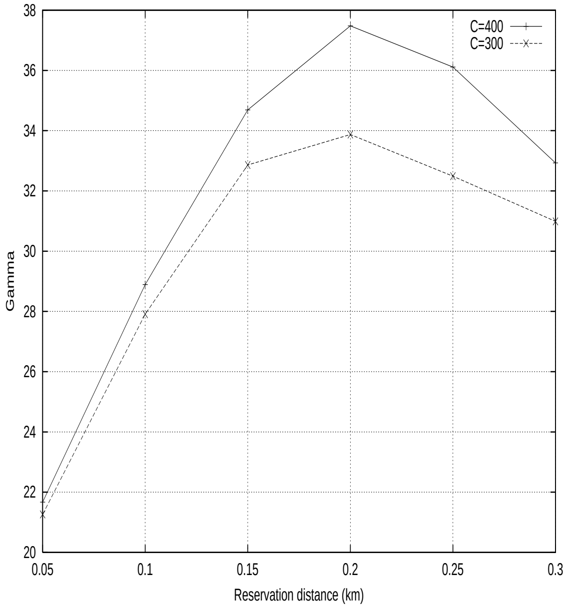

Figure 7: Impact of reservation distance on γ.

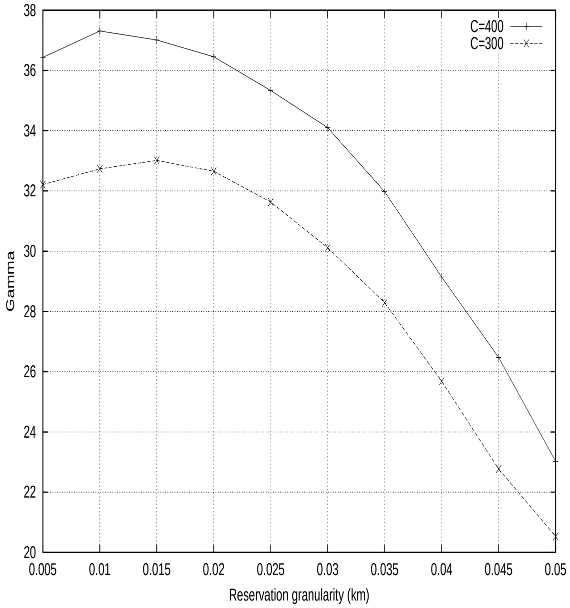

Figure 8: Impact of reservation granularity on γ.

Figure 7 shows the dependence of γ on the reservation distance, r_d. It should be pointed out that the region of the figure where $r_d > 0.3$ is not practical, since the network should not start reserving channels for a profiled user when that user is half a cell radius away from the cell boundary. Hence, we have not shown results for $r_d > 0.3$ in our figures. We observe that the there is an optimal value of r_d (at $r_d = 0.2$ km) which results in the maximum improvement in dropping probability. When r_d is small (e.g., at $r_d = 0.05$ km), the reservation attempts are made too close to the cell boundary. Hence, there is not enough time to recover from a failed reservation attempt before the handoff occurs. Therefore, the dropping probabilities of non-profiled and profiled users differ by a small margin resulting in a small γ. When r_d is too large (e.g., at 0.25 km and beyond), the channel-holding time of profiled users is inflated by a large amount which causes the overall load in a cell to increase. This results in large dropping probabilities for both profiled and non-profiled users. Thus, we observe a small improvement in the dropping probability.

Figure 8 shows the variation in γ with respect to the reservation granularity, r_g. Again, we observe an optimal value of r_g ($r_g = 0.01$ for $C = 400$) which results in the maximum improvement in dropping probability. The reason for this optimality is very similar to the one presented above. If we keep r_g small, the network makes a large number of reservation attempts on behalf of the profiled user. Though this should improve the dropping probability of profiled users, a very small value of r_g results in higher load to a cell, and hence a large dropping probability for non-profiled as well as profiled users. This results in a small γ as can be seen in the left region of the Figure 8. When r_g is large, there are not enough reservation re-attempts for profiled users. Hence, there is very little difference in dropping probabilities between non-profiled and profiled users.

Figure 9 shows γ as the fraction of profiled users (p) is changed from 0.1 to 0.9, keeping the new-call arrival rate constant at 0.025 calls/sec. We observe that the improvement increases with p till around $p = 0.3$ to $p = 0.5$, and then decreases.

7.1 Mathematical Model

To understand the problem better, and to quickly analyze the approximate benefits of our profile-based scheme, we have developed a simple analytical model for our network [25]. Figure 10 shows the results obtained through our mathematical analysis, for the default parameters shown in Table 1. The analysis shows a similar trend in γ as the simulation, including the peaks. Moreover, the analysis runs much faster, and can be used to obtain an approximate value of the improvements in dropping probability for large networks. The discrepancy between analysis and simulation can be attributed to the approximations we had made for analytical tractability. They include: (1) approximating a finite user population by an infinite one, (2) assuming that users are distributed uniformly over the network instead of being concentrated in hot-spots, (3) modeling user mobility as uniform fluid-flow with equal probability of handoff to the neighboring cells, and (4) approximating handoff calls arriving into a cell by a Poisson process.

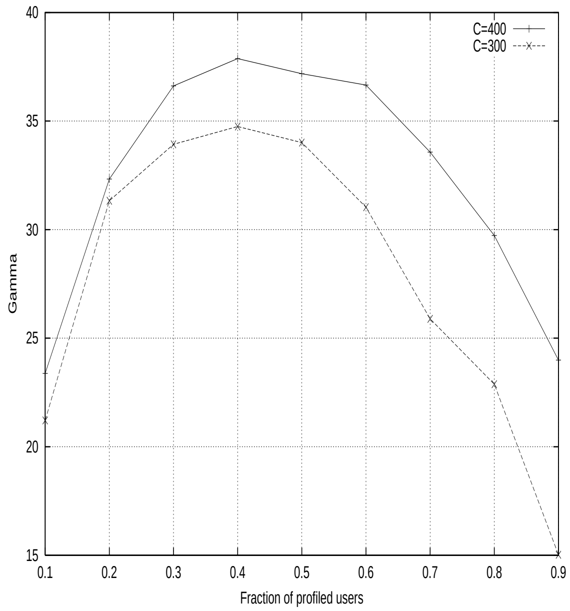

Figure 9: γ for varying fraction of profiled users.

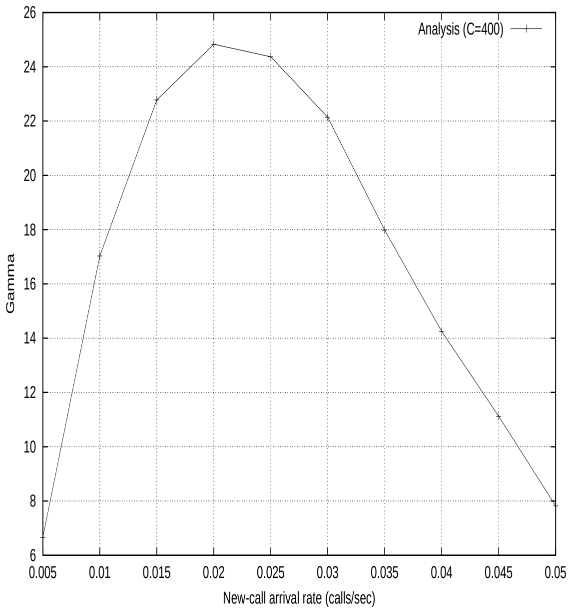

Figure 10: Mathematical analysis for γ.

7.2 Comparison with Two-Class Scheme

In this subsection, we present a comparison of the scheme proposed in [11, 12] with our scheme. Henceforth, we shall refer to this other scheme as the Two-Class (TC) scheme, and our scheme as the PARMA scheme. The TC scheme divides the number of available channels in a cell into two groups. Hence, G channels out of m are reserved beforehand for use by profiled users only. The remaining $m - G$ channels are used by all new-call arrivals and the handoffs of non-profiled users. The $m - G$ channels could also be used by the profiled users in the case when all G channels have already been occupied. Therefore, in the TC scheme, the following algorithm is followed (shown in Figure 11).

- On a new-call arrival or a handoff, check whether it is a non-profiled user. If yes, try to set up the call on one of the $m - G$ channels. If all the $m - G$ channels are busy, the call gets blocked.

- If the call is a new-call or a handoff from a profiled user, try to set up the call on one of the G channels. If all G channels are busy, try to set up the call on the remaining $m - G$ channels. If all channels are busy, the call gets blocked.

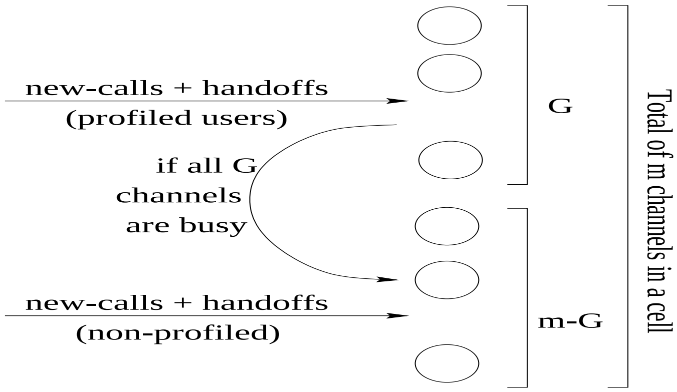

Figure 11: The Two-Class profile algorithm.

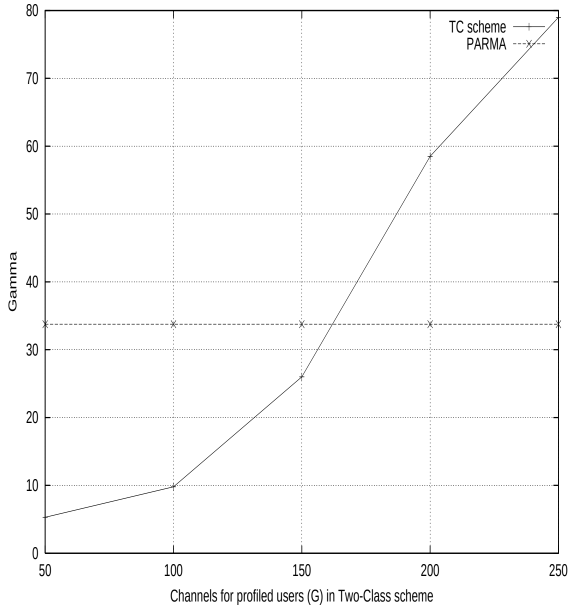

Figure 12: Comparison of γ for the two schemes.

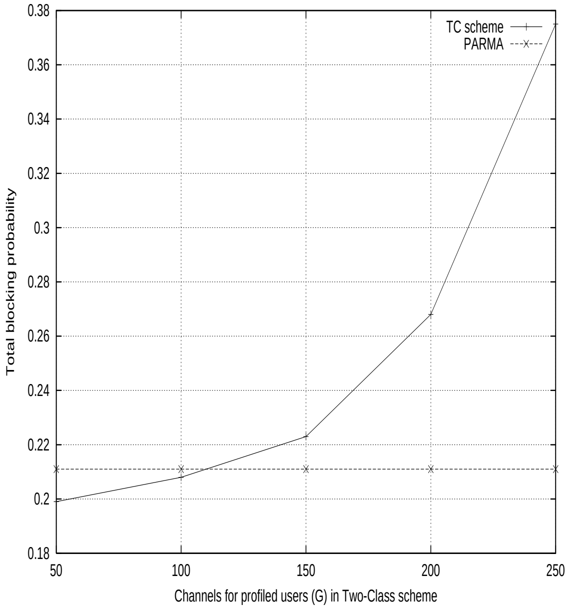

Figure 13: Comparison of total blocking probability for the two schemes.

In Figure 12, we compare the improvement in dropping probabilities obtained through both schemes. To establish a common comparison platform, we fix the new-call arrival rate to the network at 0.02 calls/sec, and the total number of channels to be $C = 300$. All other simulation parameters employ the default values described in Section 6, except G for the TC scheme, and r_d and r_g for the PARMA scheme. It should be noted that we use the network architecture defined in Section 6 for both the schemes, while the authors of the TC scheme have used a different network model.

In Figure 12, we plot the best γ we could obtain for the above-mentioned network parameters, by varying r_d and r_g for the PARMA scheme (shown as the straight line in the figure). For the TC scheme, we vary the number of reserved channels, G, and plot the improvement in dropping probability. In Figure 13, we plot the total blocking probability for the two schemes for the same operating parameters. By comparing these two figures, we make the following observations.

For a given total blocking probability, the PARMA scheme produces a larger improvement in dropping probability. This can be explained intuitively using queuing theory. The PARMA scheme can be roughly modeled as a single finite-buffer, multiple-server queue (or an M/M/m/m queue), while the TC scheme can be modeled as a multiple-server queue with two finite buffers – one each for profiled and non-profiled users. The two-buffer scheme performs less "load sharing" than the single-buffer scheme, i.e., it is more probable in the two-buffer scheme that a server is idle when another is heavily loaded. On the other hand, we can make γ very large in the TC scheme by increasing G, at the cost of increased blocking probability of non-profiled users.

8 Related Work

The idea of user profiles has been proposed in the literature in the context of the mobility management for wireless and mobile networks [26, 27, 28, 29]. These studies have focused on user tracking and location issues. The authors in [30] propose a shadow cluster scheme for estimating future resource requirements in a set of cells which a user is likely to visit in the future, but the work does not state how to determine the shadow cluster in a real network. User mobility and path prediction have been studied in [31, 32, 33, 34].

There have been a number of recent studies in the area of providing QoS to mobile users, especially for multimedia traffic [35, 36, 37, 38, 5, 39, 11, 12, 40, 17, 41].

The study in [36] proposes a call-admission-control algorithm for QoS provisioning in multimedia wireless networks. The work studies an adaptive resource-sharing policy among real-time and non-real-time traffic, where real-time traffic has preemptive priority over non-real-time traffic. The authors in [37] proposed a bandwidth-reservation scheme to guarantee QoS to multimedia traffic based on users' minimum resource-requirement specifications. Bandwidth was reserved in neighboring cells irrespective of the final destination of the mobile user. A cellular network consisting of multiple classes of users with varying bandwidth requirements has been examined in [38]. This work examined network performance with quotas set for different resources and a higher priority set for handoff calls over new calls. At the application layer, for wireless web traffic, [17] proposes a cache relocation scheme based on a mobile user's path profile to enhance the user's browsing experience within a cellular network.

The concept of graceful degradation of services in the network to support an increased load was proposed in [39]. A user can supply a loss profile, i.e., the amount of service degradation each application is willing to accept, which is used by the network to admit a larger number of users. In [40], the authors have proposed an integrated framework for QoS provisioning at the radio-link layer in packet-switched cellular networks for real-time and non-real-time traffic. The approach combines channel reservation, bandwidth compaction, and bandwidth degradation to provide differential treatment and guarantee QoS.

The two-class scheme [11, 12] mentioned in Section 7.2 proposes a channel-partitioning scheme, where one set of channels is used exclusively by the higher-priority profiled users, while the other set can be used by regular (non-profiled) users as well as profiled users. The authors provide an algorithm which can be used to reserve channels in handoff target cells. The study in [41] proposes a new user-path-prediction technique for mobile users and utilizes this algorithm to provide better QoS to all users in a cellular network.

While most previous work examined channel reservation, path prediction and call admission control [42, 43, 44] as separate issues in a cellular network, some recent studies [40, 17, 41] have attempted to combine these concepts in a unified framework. Out of this body of work, the authors of [11] and [41] study the problem of channel reservation in the context of circuit-switched cellular networks consisting of voice traffic. In [11, 12], the authors present a static partitioning of the available channels, while [41] studies channel reservation only for a single class of mobile users.

Most studies have utilized fixed allocation of a set of resources for profiled users. Such approaches would lead to higher QoS at the cost of low resource utilization. In our present work, we have studied the characteristics of an on-demand resource-reservation algorithm which provides high QoS while maintaining high resource utilization.

9 Conclusion

With the continuing deployment of intelligent network components, it is becoming easier to collect and maintain accurate real-time data on the location, the velocity, and the resource requirements of a mobile user. These data can be used to develop user profiles, and they can also be aggregated to develop mobility and resource-requirement patterns of users in a region. We have made the following contributions in this work: (1) We have described the design and implementation of a scheme, called PARMA, which utilizes user profiles to provide better QoS to mobile users in a wireless network. Specific implementation details have been proposed for the 3GPP network architecture, though the concepts would be broadly applicable to most wireless network architectures. (2) There are numerous challenges and design issues in implementing such a scheme for the next-generation wireless networks, and we attempt to resolve some of these design issues. (3) Through detailed simulation, we have studied the benefit of user profiles in improving the QoS of cellular customers. We have studied a resource-allocation scheme using the concept of reservation distance and reservation granularity. We have shown that this concept can produce significant improvement in dropping probability of profiled users over their non-profiled counterparts. We showed that there are optimal values of the reservation distance and the reservation granularity parameters which result in maximal improvement in dropping probability.

References

[1] UMTS Forum, http://www.umts-forum.org, *Universal Mobile Telecommunications System.*

[2] International Telecommunication Union, http://www.imt-2000.org, *International Mobile Telecommunication - 2000.*

[3] 3rd Generation Partnership Project, *Network Architecture (T1TRQ3GPP 23.002-330)*, Standards Committee T1 Telecommunications, version 3.3.0 edition, 2000.

[4] V. O. K. Li and X. Qiu, "Personal communication systems (PCS)," *Proceedings of the IEEE*, vol. 83, no. 9, pp. 1210–1243, September 1995.

[5] M. Naghshineh, M. Schwartz, and A.S. Acampora, "Issues in wireless access broadband networks," *Wireless Information Networks, Resource Management, and Mobile Data, edited by J.M. Holtzman*, 1996, Kluwer Academic Publishers.

[6] Y. Fang and I. Chlamtac, "A new mobility model and its application in the channel holding time characterization in pcs networks," in *Proc., IEEE INFOCOM*, New York, NY, March 1999, IEEE, vol. 1, pp. 20–27.

[7] J.G. Markoulidakis, G.L. Lyberopoulos, and M.E. Anagnostou, "Traffic model for Third Generation cellular mobile telecommunications systems," *ACM/Baltzer Wireless Networks*, vol. 4, pp. 389–400, August 1998.

[8] I. F. Akyildiz, Y.-B. Lin, W.-R. Lai, and R.-J. Chen, "A new random walk model for PCS networks," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 7, pp. 1254–1260, July 2000.

[9] J.S.M. Ho and I.F. Akyildiz, "Dynamic mobile user location update for wireless PCS networks," *ACM/Baltzer Wireless Networks*, vol. 1, no. 2, pp. 187–196, December 1995.

[10] B. Jabbari and W. Fuhrmann, "Teletraffic modelling and analysis of flexible hierarchical cellular networks with speed sensitive handoff strategy," *IEEE Journal on Selected Areas in Communications*, vol. 15, no. 8, pp. 1539–1548, October 1997.

[11] A. N. Rudrapatna, D. P. Agrawal, P. Agrawal, and C. Giardina, "User profile-driven resource allocation for cellular networks," in *Proc., IEEE International Conference on Personal Wireless Communications*, Mumbai, India, December 1997, pp. 178–183.

[12] A. N. Rudrapatna and C. Giardina, "Channel occupancy and network utilization implications for a profile-driven resource allocation scheme in cellular networks," in *Proc., of GLOBECOM*, Sydney, Australia, November 1998, vol. 4, pp. 2007–2012.

[13] M-A. Dru and S. Saada, "Location-based mobile services: The essentials," Tech. Rep., Alcatel Telecommunications Review, 1st Quarter 2001.

[14] Nokia, http://www.nokia.com, *Mobile Location Services*, 2001.

[15] K. L. Yueng and S. Nanda, "Channel manangement in microcell/macrocell cellular radio systems," *IEEE Transactions on Vehicular Technology*, vol. 45, no. 4, pp. 601–612, November 1996.

[16] E. W. Knightly and N. B. Shroff, "Admission control for statistical QoS: theory and practice," *IEEE Network*, vol. 13, no. 2, pp. 20–29, March-April 1999.

[17] S. Hadjiefthymiades and L. Merakos, "Using proxy cache relocation to accelerate web browsing in wireless/mobile communications," in *Proc., World Wide Web Conference*, Hong Kong, May 2001, vol. 10.

[18] G. P. Pollini, "Trends in handover design," *IEEE Communications*, vol. 34, no. 3, pp. 82–90, March 1996.

[19] S.A. El-Dolil, W.-C. Wong, and R. Steele, "Teletraffic performance of highway microcells with overlay macrocells," *IEEE Journal on Selected Areas in Communications*, vol. 7, no. 1, pp. 71–78, January 1989.

[20] H. Furukawa and Y. Akaiwa, "A microcell overlaid with umbrella cell system," in *Proc., IEEE Vehicular Technology Conference*, Stockholm, Sweden, June 1994, vol. 3, pp. 1455–1459.

[21] Y-B. Lin, L. F. Chang, and A. Noerpel, "Modeling hierarchical microcell/macrocell PCS architecture," in *Proc., IEEE International Conference on Communications*, Seattle, WA, June 1995, vol. 1, pp. 405–409.

[22] V. Pandey, D. Ghosal, and B. Mukherjee, "Performance issues in two-tier cellular networks," in *IEEE International Conference on Personal Wireless Communication*, Jaipur, India, February 1999, pp. 374–378.

[23] I. Katzela and M. Naghshineh, "Channel assignment schemes for cellular mobile telecommunication systems: A comprehensive survey," *IEEE Personal Communications*, pp. 10–31, June 1996.

[24] A. D. May, *Traffic Flow Fundamentals*, Prentice Hall, 1990.

[25] V. Pandey, *Resource Management in Integrated Wireless Networks*, Ph.D. thesis, University of California, Davis, 2002.

[26] S. Tabbane, "An alternative strategy for location tracking," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 5, pp. 880–892, June 1995.

[27] Y-B. Lin, "Reducing location update cost in a PCS network," *IEEE/ACM Transaction on Networking*, vol. 5, no. 1, pp. 25–33, February 1997.

[28] S. Tabbane, "Location management methods for third-generation mobile systems," *IEEE Communications*, vol. 35, no. 8, pp. 72–85, August 1997.

[29] C. H. Rokitansky, "Knowledge-based routing strategies for large mobile networks with rapidly changing topology," in *Proceedings of International Conference on Computers and Communication*, New Delhi, India, November 1990, pp. 541–550.

[30] D.A. Levine, I.F. Akyildiz, and M. Naghshineh, "A resource estimation and call admission algorithm for wireless multimedia networks using the shadow cluster concept," *IEEE/ACM Transactions on Networking*, vol. 5, no. 1, pp. 1–12, February 1997.

[31] S. Lu and V. Bharghavan, "Adaptive resource management algorithms for indoor mobile computing environments," in *Proc., ACM SIGCOMM*, Stanford, CA, September 1996, pp. 231–242.

[32] S. Choi and K. G. Kin, "Predictive and adaptive bandwidth reservation for handoffs in QoS-sensitive cellular networks," in *Proc., ACM SIGCOMM*, Vancouver, BC, September 1998, pp. 155–166.

[33] M. H. Chiu and M. A. Bassiouni, "Predictive schemes for handoff prioritization in cellular networks based on mobile positioning," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 3, pp. 510–522, March 2000.

[34] A. Bhattacharya and S. K. Das, "Lezi-update: An information theoretic approach for personal mobility tracking in PCS networks," *ACM/Baltzer Wireless Networks*, vol. 8, no. 2-3, pp. 121–135, March-May 2002.

[35] S. Papavassiliou, L. Tassiulas, and P. Tandon, "Meeting QoS requirements in a cellular network with reuse partitioning," in *Proc., IEEE INFOCOM*, Toronto, Canada, June 1994, IEEE, vol. 1, pp. 4–12.

[36] A.S. Acampora and M. Naghshineh, "QoS provisioning in microcellular networks supporting multimedia traffic," in *Proc., IEEE INFOCOM*, Boston, MA, April 1995, IEEE, vol. 3, pp. 1075–1084.

[37] C. Oliveira, J. B. Kim, and T. Suda, "Quality of Service guarantee in high-speed multimedia wireless networks," in *Proc., IEEE International Communications Conference*, Dallas, Texas, June 1996, IEEE, vol. 2, pp. 728–734.

[38] S. Rappaport and C. Purzynski, "Prioritized resource assignment for mobile cellular communication systems with mixed services and platform types," *IEEE Transactions on Vehicular Technology*, vol. 45, no. 3, pp. 443–458, August 1996.

[39] K. Seal and S. Singh, "Loss profiles: A Quality of Service measure in mobile computing," *Journal of Wireless Computing*, vol. 2, pp. 45–61, 1996.

[40] R. Jayaram, N. K. Kakani, S. K. Das, and S. K. Sen, "A call admission and control scheme for Quality of Service (QoS) provisioning in next generation wireless networks," *ACM/Baltzer Wireless Networks*, vol. 6, no. 1, pp. 17–30, 2000.

[41] F. Yu and C. M. Leung, "Mobility-based predictive call admission control and banmdwidth reservation in wireless cellular networks," in *Proc., IEEE INFOCOM*, Anchorage, AK, April 2001, IEEE, vol. 1, pp. 518–526.

[42] E. C. Posner and R. Guerin, "Traffic policies in cellular radio that minimize blocking of handoff calls," in *Proceedings of the ITC-11*, Kyoto, Japan, September 1985.

[43] R. Ramjee, D. Towsley, and R. Nagarajan, "On optimal call admission control in cellular networks," *ACM/Baltzer Wireless Networks*, vol. 3, no. 1, pp. 29–41, March 1997.

[44] M. Naghshineh and M. Schwartz, “Distributed call admission control in mobile/wireless networks,” *IEEE Journal on Selected Areas in Communications*, vol. 14, no. 4, pp. 711–717, May 1996.